

RIDING THE WAVE

**FROM VISIONS TO ACTIONS FOSTERING
EUROPEAN RESEARCH
HLEG ON SCIENTIFIC DATA**

**- AN ACTION FROM THE EUROPEAN
COMMISSION -**



Peter Wittenburg
The Language Archive - Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

- ❑ **Group and Motivation**
- ❑ The Research Data World
- ❑ Opportunities and Challenges
- ❑ Collaborative Data Infrastructure
- ❑ Relevant Aspects
- ❑ Vision 2030
- ❑ Action Points

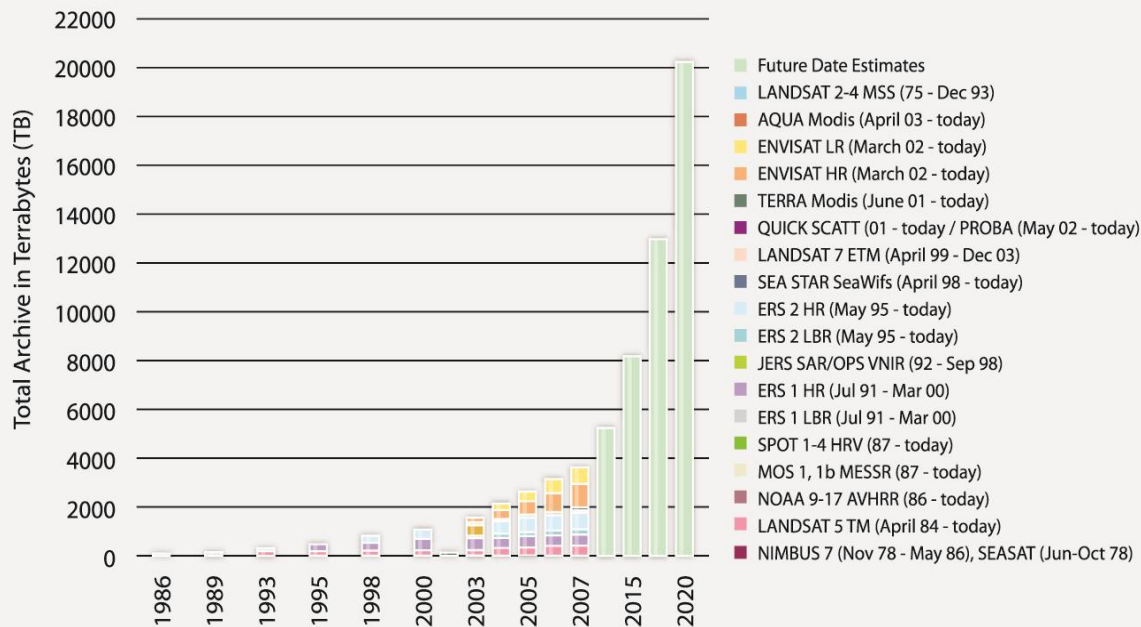
HLEG and Motivation

- EC DG InfSo invited 11 experts and 12 contributors to 6 meetings
Chair: John Wood; **Rapporteur:** David Giaretta; **Members:** Thomas Andersson ; Achim Bachem; Christoph Best ; Françoise Genova ; Diego R. Lopez; Wouter Los; Monica Marinucci; Laurent Romary; Herbert Van de Sompel ; Jens Vigen; Peter Wittenburg; **EC Representatives:** Konstantinos Glinos; Carlos Morais-Pires;
- Goals
 - come up with a vision 2030 for the management of research data as a guideline for future actions of the EC
 - discuss all relevant aspects around “data” in an unbiased manner
 - accelerate measures to take care of our data and to remain competitive
- Motivation
 - enormous increase in scale and complexity
 - not only summarize what some of us already know or are doing, but facilitate a systematic and global approach and push ahead actions
 - knowledge is power - data has a value although difficult to quantify

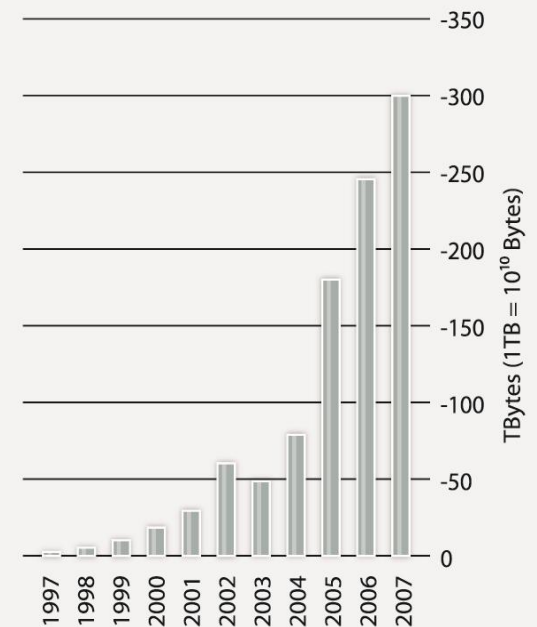
Trends are Known


“A fundamental characteristic of our age is the raising tide of data – **global, diverse, valuable and complex**. In the realm of science, this is both an **opportunity** and a **challenge**.”

Evolution of ESA's EO Data Archives between 1986-2007 and future estimates (up to 2020)



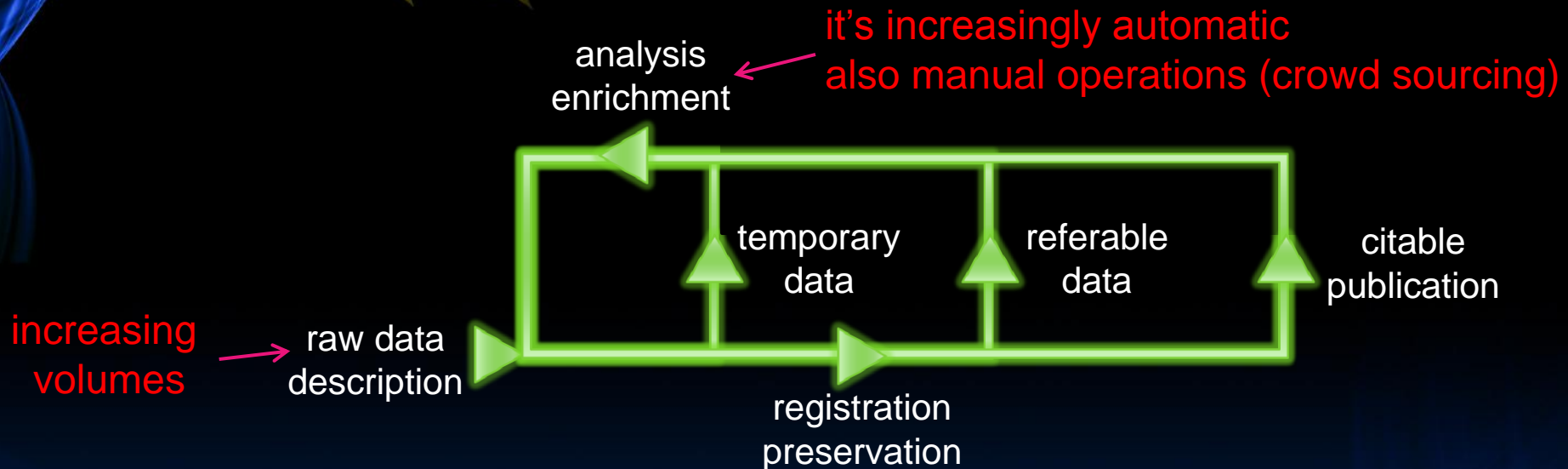
Yearly Data Creation on NICE



- 
- ❑ Group and Motivation
 - ❑ **The Research Data World**
 - ❑ Opportunities and Challenges
 - ❑ Collaborative Data Infrastructure
 - ❑ Relevant Aspects
 - ❑ Vision 2030
 - ❑ Action Points

Research Data World

- Knowledge Creation Cycle is changing - almost all disciplines



- Exabyte scale and millions of related files of different types create unseen complexity - **deal with a new quality**
- much relevant data is and will not be registered (80 % of recordings about languages and cultures are endangered)

Berman's classification

The data pyramid - a hierarchy of rising value and permanence

Digital Data Collections

Reference, nationally and internationally important, irreplaceable data collections

Key research and community data collections

Personal data collections

Increasing constituency

Increasing value

Increasing trust

Societal Value
Patrimonial Data

Community Value
Cyclic Data

Individual Value
Transient Data

decreasing risk of loss or damage

Increasing responsibility

Increasing stability

Increasing infrastructure

Respositories/ Facilities

National- and international-scale respositories, libraries, archives

"Regional" - scale libraries and targeted data archives and centers

Private respositories

Source: Adapted from Francine Berman, UC San Diego, in *Communications of the ACM*.

this is the data we need to take care of but do we know which data will be of relevance for future generations?

- some interesting aspects
 - lossless **separation of content and carrier** in the digital domain changes the world - some speak about a revolution comparable with the invention of book printing
 - **data creators are not known personally** to data users anymore - we need to solve the trust problem
 - research world is one of the **primary driver** for the data tides
 - there is no doubt: **data accessibility** changes nature, pace and direction of research
 - **diversity** in many dimensions is the dominant feature of scientific information and this will probably increase due to the inherent innovation forces
 - technology allows to **include the citizens** in different roles - also as contributors, increasing volume and complexity
 - increasing pressure towards **open access**

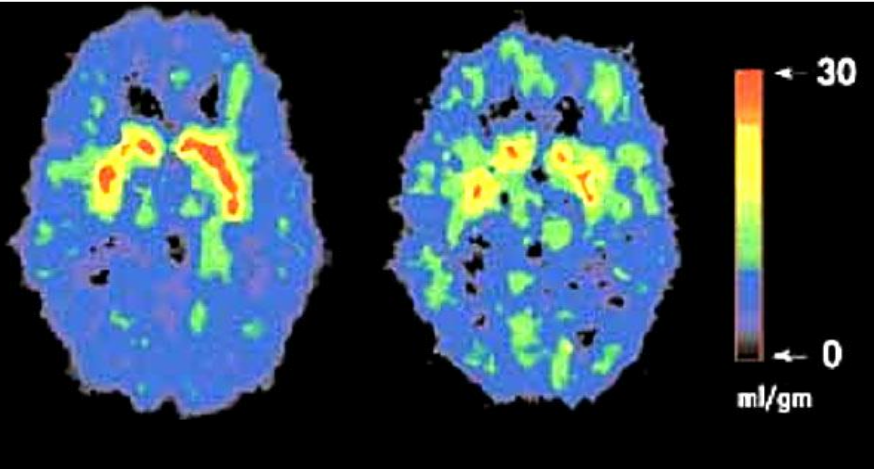
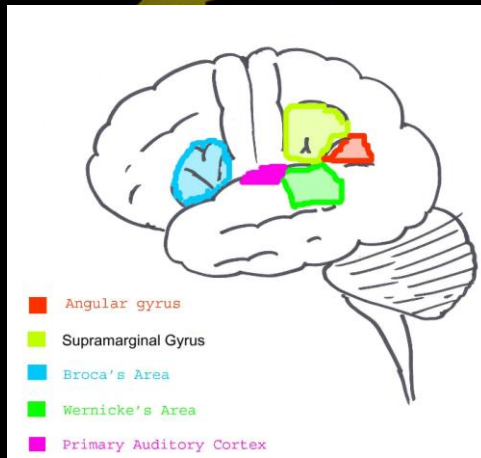
- 
- ❑ Group and Motivation
 - ❑ The Research Data World
 - ❑ **Opportunities and Challenges**
 - ❑ Collaborative Data Infrastructure
 - ❑ Relevant Aspects
 - ❑ Vision 2030
 - ❑ Action Points

Opportunities and Challenges

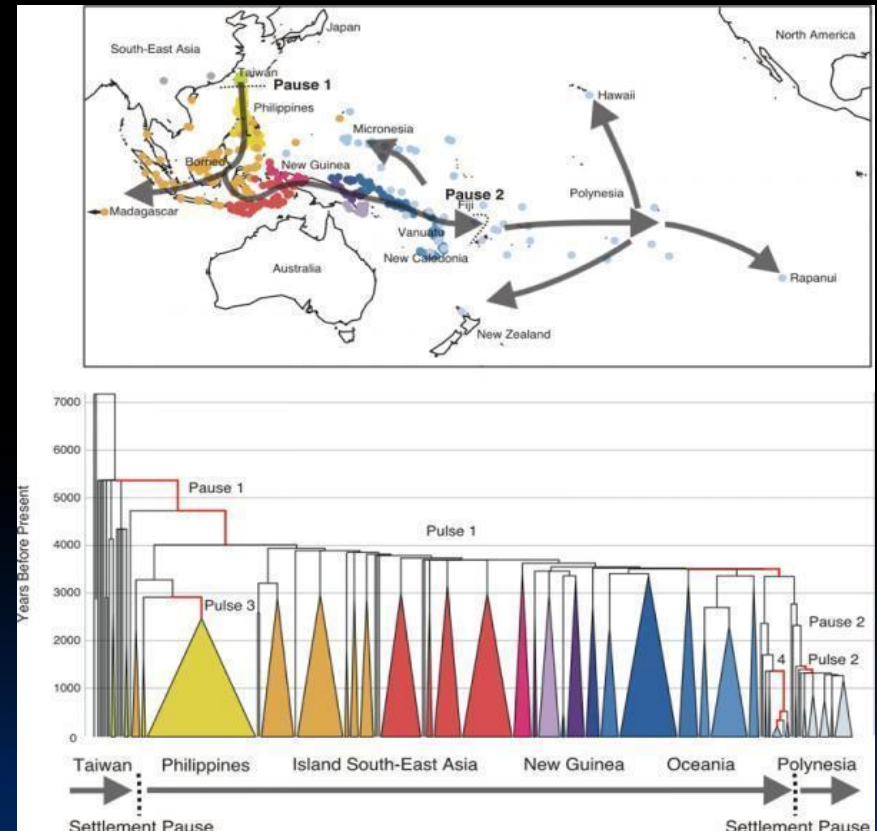
- virtual integration
 - **integrating large data** sets across disciplines and countries to create new insights
 - recombining data to **virtual collections** from different perspectives
 - sufficient data as basis for **comprehensive modeling** and understanding
 - **data intensive science**: find correlations and draw inferences not constraint by pre-assumptions - huge amounts of data not used
- tackling the **grand challenges** resulting from human activities
 - climate change, sustainable energy, stability health, etc.
 - **stability of our societies and minds** given the innovation, changes, globalization and migration
- facilitating the many “**small research questions**” driven by scientific curiosity
- relieve researchers from **data management and curation** effort (40% of knowledge workers time spent on finding and transforming data)

Opportunities and Challenges

two examples for “small research questions” from my domain:
languages and language processing



finding more about the functional
architecture of our brain



finding more about the roots of
our languages and cultures

Opportunities and Challenges

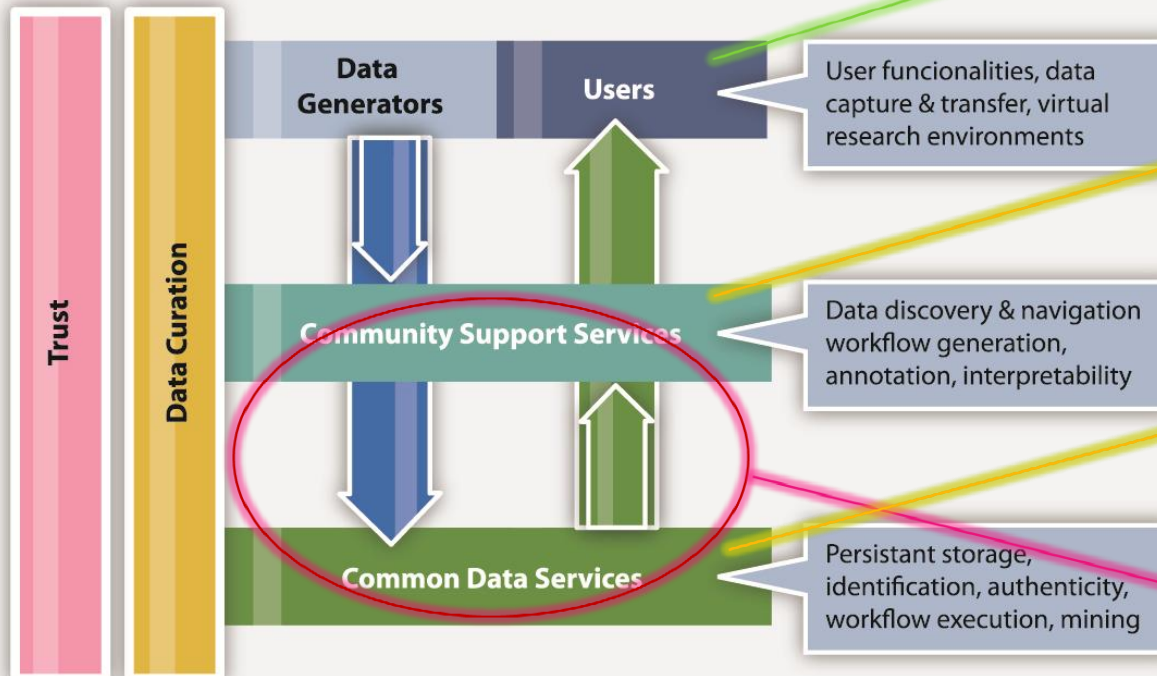
- however there are quite some hurdles to overcome
 - need to **change culture** and researchers minds to deposit data
 - need to establish **trust** at depositor's and user's side
 - trust has to do with **data quality, integrity and authenticity**
 - need to convey **context and provenance** to allow users to understand
 - need **new responsibilities** and **new mechanisms** to solve data curation, preservation, organization and granting access without ignoring security and ownership principles
- need **incentives** for researchers to deposit in proper quality so that data publication helps in career and reputation building

- 
- ❑ Group and Motivation
 - ❑ The Research Data World
 - ❑ Opportunities and Challenges
 - ❑ **Collaborative Data Infrastructure**
 - ❑ Relevant Aspects
 - ❑ Vision 2030
 - ❑ Action Points

Collaborative Data Infrastructure

- obviously we need a new layer of responsibility:
a **systematically constructed and global data infrastructure**
- some already working on data organizations - piecemeal, fragmented
- we call it a **Collaborative Data Infrastructure** open for many players and heterogeneity based on an abstract architecture and proper APIs

The Collaborative Data Infrastructure - a framework for the future



many researchers from different disciplines and with different interests

CLARIN, CESSDA, DARIAH, LifeWatch, ENES, EPOS, etc.
>40 R Infrastructures

EUDAT, OpenAIRE, D4Science, etc.

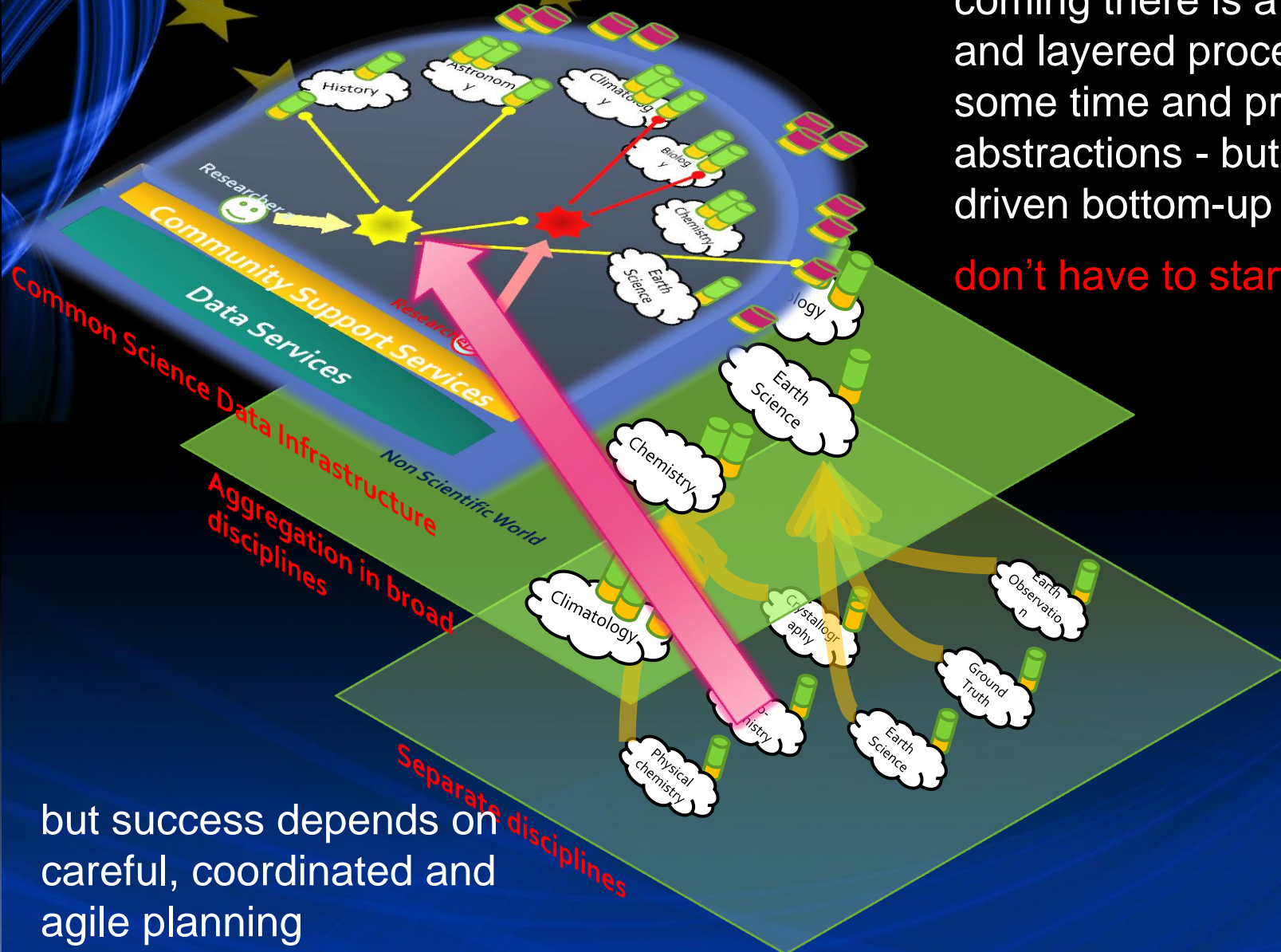
are faced with large heterogeneity
need an architecture

- speaking about abstractions
 - data **object level** vs. data **content level**
 - can put cherries, apples, potatoes, etc. all in the same container
 - the way to treat them for making food is different
 - data object architecture is common (?)
 - early Internet discussion: is **email** specific for a discipline?
 - **PID** to identify, to ensure integrity and authenticity, etc.
 - **metadata** to describe context and provenance
 - how to treat **collections** (ORE, etc.)
 - have **many instances** (copies) at various locations (preservation, etc.)
 - **interoperability** at object level is about MD, PIDs etc.
 - data content level is discipline specific
 - it's about **structure** (schemas) and encoding schemes (MPEGx, etc.)
 - it's about **semantics** (vocabularies)
 - **interoperability** is discipline and **USAGE** specific

Collaborative Data Infrastructure

coming there is a step-wise
and layered process requiring
some time and proper
abstractions - but needs to be
driven bottom-up

don't have to start from scratch



but success depends on
careful, coordinated and
agile planning

- ❑ Group and Motivation
- ❑ The Research Data World
- ❑ Opportunities and Challenges
- ❑ Collaborative Data Infrastructure
- ❑ **Relevant Aspects**
- ❑ Vision 2030
- ❑ Action Points

Relevant Aspects

- funding
 - need to understand data as a socio-economic treasure in a competitive domain - at the end research is about **global competition**
 - need proper **business models** - who is paying, which data is free, etc.
 - **governments** will have to reserve funds for data management
- quality and impact
 - need to measure quality and impact, which **metrics** are meaningful
 - need to **reward** contributors but how?
- management/curation skills
 - need a new type of experts: **data scientists**
- power researchers
 - resulting CDI will be complex as the data world will be
 - need to **educate and train** a new generation of power users
- ecology
 - uncontrolled **copying** of data sets is not ecological
 - need to take care of **green computing** principles

- ❑ Group and Motivation
- ❑ The Research Data World
- ❑ Opportunities and Challenges
- ❑ Collaborative Data Infrastructure
- ❑ Relevant Aspects
- ❑ **Vision 2030**
- ❑ Action Points

Vision 2030

All **stakeholders**, from scientists to national authorities to general public are aware of the critical importance of preserving and sharing reliable data produced during the scientific process.

Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data and they can evaluate the degree to which the data can be trusted.

Producers of data benefit from opening it to broad access and prefer to deposit their data with confidence in reliable repositories. A framework of repositories work to international standards, to ensure they are trustworthy.

Vision 2030

Public funding rises, because funding bodies have confidence that their investments in research are paying back extra dividends to society, through increased use and re-use of data.

The innovative power of **industry and enterprise** is harnessed by clear and efficient arrangements for exchange of data.

The **public** has access and can make creative use of the huge amount of data available; it can also contribute to the data store and enrich it.

Policy makers can make decisions based on solid evidence, and can monitor the impacts of these decisions.

Global governance promotes international trust and interoperability.

- ❑ Group and Motivation
- ❑ The Research Data World
- ❑ Opportunities and Challenges
- ❑ Collaborative Data Infrastructure
- ❑ Relevant Aspects
- ❑ Vision 2030
- ❑ **Action Points**

Action Points

- HLEG requests
 - need a CDI initiative
 - earmark **additional funds**
 - develop new ways to measure data **value** and reward researchers
 - train a new generation of **data scientists** and broaden understanding
 - think **green**
 - establish a **high level coordination group**
 - asking for **global collaboration**
- Recent Actions
 - **EUDAT** received grant to work on CDI
 - together with OpenAIRE work on establishing a **Data Access and Interoperability Task Force**
 - EC call for collaboration with US projects
 - DAITF preparation workshop in Copenhagen (20/21. March)
 - new e-IRG document on Data Management

- Recent EC Statements (Communication Dec 2011)
 - strong statement for **open and aggregated (meta) data**
 - EC will invest in **data infrastructures** which are based on **distributed and participatory architectures**
 - robust networks with national, regional and domain specific hubs
 - open data portals and platforms (European Data Portal)
 - supporting research & innovation for re-using, re-purposing data
 - changing the **legislation** (default = re-usage)
 - lowering **fees** to dissemination costs



Thanks for your attention.